Assessing QSAR Limitations – A Regulatory Perspective

Weida Tong[1,*], Huixiao Hong[2], Qian Xie[2], Leming Shi[1], Hong Fang[2] and Roger Perkins[2]

[1]*Center for Toxicoinformatics, National Center for Toxicological Research (NCTR), Jefferson, AR 72079, USA*

[2]*Division of Bioinformatics, Z-Tech Inc., Jefferson, AR 72079, USA*

Abstract: Wider acceptance of QSARs would result in a constellation of benefits and savings to both private and public sectors. For this to occur, particularly in regulatory applications, a model's limitations need to be identified. We define a model's limitations as encompassing assessment of overall prediction accuracy, applicability domain and chance correlation. A general guideline is presented in this review for assessing a model's limitations with emphasis on and examples of application with consensus modeling methods. More specifically, we discuss the commonalities and differences between external validation and cross-validation for assessing a model's limitations. We illustrate two common ways of assessing overall prediction accuracy, depending on whether or not the intended application domain is predefined. Since even a high quality model will have different confidence in accuracy for predicting different chemicals, we further demonstrate using the novel Decision Forest consensus modeling method a means to determine prediction confidence (i.e., certainty for an individual chemical's prediction) and domain extrapolation (i.e., the prediction accuracy for a chemical that is outside the chemistry space defined by the training chemicals). We show that prediction confidence and domain extrapolation are related measures that together determine the applicability domain of a model, and that prediction confidence is the more important measure. Lastly, the importance of assessing chance correlation is emphasized, and illustrated with several examples of models having a high degree of chance correlations despite cross-validation indicating high prediction accuracy. Generally, a dataset with a skewed distribution, small data size and/or low signal/noise ratio tends to produce a model with high chance correlation.

We conclude that it is imperative to assess all three aspects (i.e., overall accuracy, applicability domain and chance correlation) of a model for the regulatory acceptance of QSARs.

Keywords: SAR/QSAR, model limitation, model uncertainty, applicability domain, model validation, chance correlation, decision forest, consensus modeling.

INTRODUCTION

Quantitative structure-activity relationships (QSARs) have been widely used in the pharmaceutical industry, primarily for lead discovery and optimization. QSARs have also been employed in toxicology [1,2] and regulation [3-5] and have been particularly cost effective for prioritizing untested chemicals for more extensive and costly experimental evaluation. However, for QSARs to receive wider acceptance by regulatory communities, their limitations of use needs to be identified [6-9].

Obtaining a good quality QSAR model depends on many factors, such as the quality of biological data and the choice of descriptors and statistical methods. Given the technological advances and broader availability of various statistical methods and types of descriptors, it is now relatively easy and straightforward to develop a statistically sound model. However, methods for quantifying QSAR limitations of usage have not been addressed adequately and represent a current challenge to those working in the field [10]. The limitation of a QSAR model's applicability can be characterized in three distinct areas: (1) overall quality; (2) applicability domain; and (3) chance correlation.

The importance of validation has been generally acknowledged, and most QSAR models in the literature are validated either by cross-validation or external test sets [11,12]. *Model validation* for classification models is typically specified by statistical quality measures of *overall quality* such as sensitivity, specificity, false positives, false negatives and overall prediction. Unfortunately, it is not typical and often not possible to specify accuracy and prediction confidence for individual unknown chemicals, specifically those unknown chemicals with structures requiring that the model extrapolate beyond the chemistry space determined by the training set. Differing from overall model validation, the assessment of *applicability domain* involves determining a model's *confidence level in each prediction*. Model validation and applicability domain are separate assessments addressing model predictivity from distinct but related perspectives. Assessing chance correlation is another measure of QSAR quality that is often not provided with a QSAR model. Assessment of chance correlation intends to determine whether a valid model can be developed in the first place. This usually accomplished by averaging many models where activity classification (dependent variable) of a training set are randomly shuffled. Generally speaking, training data of small number, with skewed activity distributions and low signal to noise ratio

*Address correspondence to this author at the Center for Toxicoinformatics, NCTR, 3900 NCTR Road, HFT 020, Jefferson, AR 72079, USA; Tel: (870) 543-7142; Fax: (870) 543-7662; E-mail: Wtong@nctr.fda.gov

The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration

1573-4099/05 $50.00+.00 © 2005 Bentham Science Publishers Ltd.

are most vulnerable to chance solutions, especially with a larger number of descriptors.

Among the two major types of QSARs, classification- and regression-based, this review primarily focuses on quantitatively assessing the limitations of a classification model, with examples presented for two-class SAR models. Assessing limitations of regression-based QSARs would be theoretically similar, but differ in implementation. The usual QSAR approach is to develop a single model with as large and diverse of a training set as possible. This approach inherently results in an applicability domain that is difficult to determine and that is highly constrained by the training set. Here we present an alternative approach based on consensus of predictions of multiple models where the applicability domain is readily determinable.

This review will first present an overview of the methodological approaches to consensus modeling. Next, we will discuss the commonalities and differences between cross-validation and external validation in assessing a QSAR's limitations. Several examples will be given to illustrate how to assess a model's limitations for a novel consensus method, Decision Forest (DF), developed in our lab, including model validation, applicability domain and chance correlation. The emphasis of the review will be our current thinking and direction in QSARs with respect to the regulatory acceptance of QSARs.

CONSENSUS MODELING

Consensus modeling has been investigated for many years in the field of statistics as a means of combining multiple individual models to produce better single predictions [13]. A thorough review of this subject can be found in a number of papers [14-16]. The critical and implicit assumption in consensus modeling is that multiple models will effectively identify and encode more aspects of the SAR relationship than will a single model [17]. The corollaries are that combining several identical models produces no gain, and benefits of combining can only be realized if individual models give different predictions. In other words, benefits of combining are only expected if separate models encode differing aspects of how activity depends on structure. An ideal combined system should consist of several accurate but independent models. More recently, we also found that the information gained from combining models provides valuable indication in assessing prediction confidence for individual chemicals, which is usually difficult to obtain from a single SAR model [17].

In consensus modeling, individual models can be developed based on either the same or different methods. For example, we developed a screening procedure using four phases of separate models based on different methods for prioritizing potential environmental endocrine disruptors [18,19]. In Phase II, we combined 11 models, i.e., three structural alerts, seven pharmacophores and one Decision Tree (DT) model. The combined results dramatically reduced false negatives, which is important for regulatory applications. Alternatively, the individual models can be developed using the same modeling methods, such as artificial neural networks [20-22] and DT [23,24]. Combining models derived from the same method can be effective in reducing noise-induced error in individual models.

There are many approaches to consensus SAR modeling. Among the simplest is developing individual models using differing chemicals that are randomly selected from the entire original training dataset [25], as illustrated in Fig. **1**. Recently, we found that even combining ten models based on ten datasets from a 10-fold cross-validation procedure enhanced the overall performance of prediction [26]. Alternatively, the training set can be generated using more robust statistical "resampling" approaches, such as Bagging [27] or Boosting [28] (Fig. **2**). Both methods have been reported to improve predictive accuracy. However, it has also been demonstrated that when Boosting heavily weighs incorrect predictions during individual model development, there is inherent risk of over fitting noise associated with the SAR data, resulting in a worse prediction from an ensemble model [28].

The aforementioned resampling approaches use only a portion of the dataset for constructing the individual models. Care must be taken because using a substantial portion of datasets (e.g., 90%) tends to result in individual models that are highly correlated, whereas using a less substantial portion of datasets (e.g., 70%) tends to result in individual models of lower quality. Either highly correlated or lower quality individual models can reduce the benefit obtained from combining that might otherwise be realized. Moreover, since each chemical in a dataset encodes some SAR information, reducing the number of chemicals in a training set for model construction will weaken most individual models' predictive accuracy. Logically, it follows that reducing the number of chemicals also reduces the improvement in a combining system gained by the resampling approach. Having too few chemicals in the training set is too frequent a problem in SAR modeling that compromises both individual and consensus models.

Alternatively, multiple models can be developed using different sets of descriptors [29]. One popular DT-based consensus method (consensus tree method), random forests, has been demonstrated to be more robust than a Boosting method [30]. In this method, a small number (subset) of descriptors is randomly selected from the original descriptor pool in every split for growing a tree, and a descriptor in the subset giving the best split is chosen for splitting. Usually, this method needs to grow a large number of individual trees (>400) for convergence and could generate many correlated trees that increase bias in prediction. To keep correlation low, each tree is grown on a boostrap sample (resampling) of the training set in the current version of random forest. However, in a recent example of applying this method for classification of naïve *in vitro* drug treatment sample based on gene expression data showed reduced prediction accuracy of random forests (83.3%) compared to DT (88.9%)[31].

It is important to note that the aforementioned techniques rely on *random* selection of either samples or descriptors to produce individual models. In each repeat, the individual models of the ensemble are different, making the biological interpretation of the ensemble less straightforward. Additionally, since only a portion of the original training chemicals is used for developing individual models, all SAR information in the training set is not incorporated,

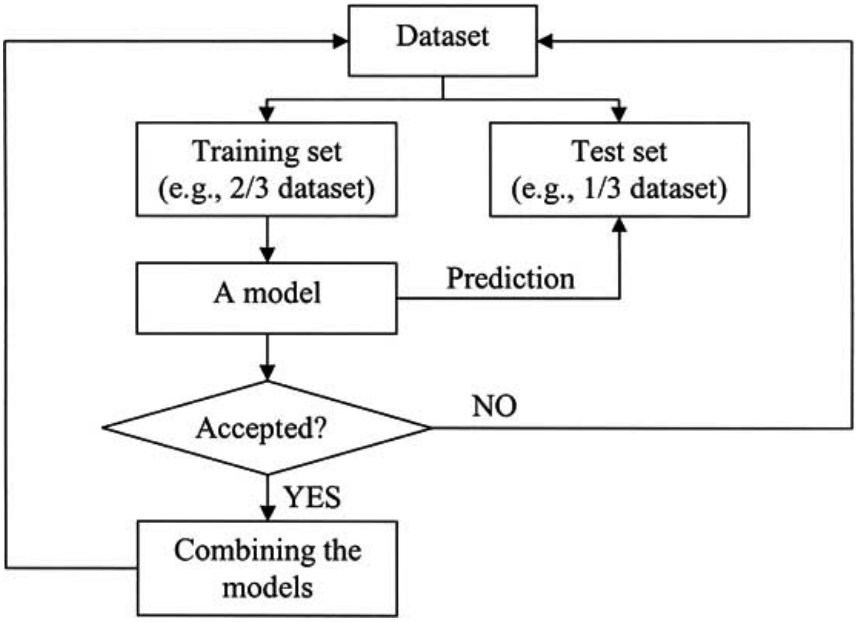


Fig. (1). A resampling method. A dataset is first randomly divided into two sets, e.g., 2/3 for training and 1/3 for testing. A model developed with the training set is accepted if it gives satisfactory predictions for the testing set. A set of predictive models is generated by repeating the procedure, and the predictions of these models are then combined when predicting a new chemical.

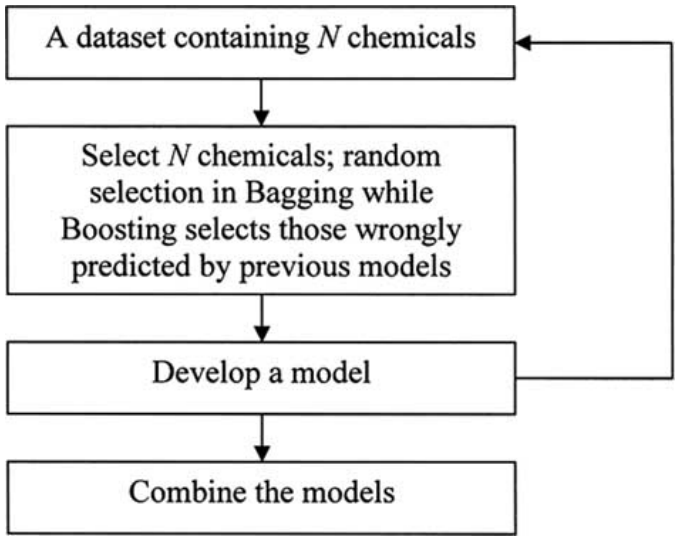


Fig. (2). Consensus modeling based on Bagging and Boosting. Bagging is a "bootstrap" ensemble method by which each model is developed on a training set that is generated by randomly selecting chemicals from the original dataset. In the selection process, some chemicals may be repeated more than once while others may be left out so that the training set is the same size as the original dataset. In Boosting, the training set for each model is also the same size as the original dataset. However, each training set is determined based on the performance of the earlier model(s); for the next training set, chemicals that were incorrectly predicted by the previous model are chosen more often than chemicals that were correctly predicted.

possibly reducing the full benefits otherwise attainable when combining multiple models.

Recently, we reported a novel DT-based consensus modeling method, named Decision Forest (DF) [17] that emphasizes the combining of *heterogeneous* yet *comparable* trees in order to better capture the association of structure and biological activity. The heterogeneity requirement assures that each tree uniquely contributes to the combined prediction; whereas the quality comparability requirement assures that each tree equally contributes to the combined prediction. Since a certain degree of noise is always present in biological data, optimizing a tree inherently risks over fitting the noise. DF attempts to minimize over fitting by maximizing the difference among individual trees, which could result in cancellation of some random noise when trees are combined. As depicted in Fig. **3**, the maximum difference of trees is achieved by constructing each individual tree using a distinct set of descriptors. There are three derived benefits associated with DF compared with other similar consensus modeling methods: (1) since the

difference in individual trees is maximized, a best ensemble is usually realized by combining only a few trees (i.e., 4 or 5), which consequentially reduces computational expense ; (2) since DF is entirely reproducible, the SAR relationships are constant in their interpretability for biological relevance; and (3) since all chemicals are included in individual tree development, the SAR information in the original dataset is fully appreciated in the combining process.

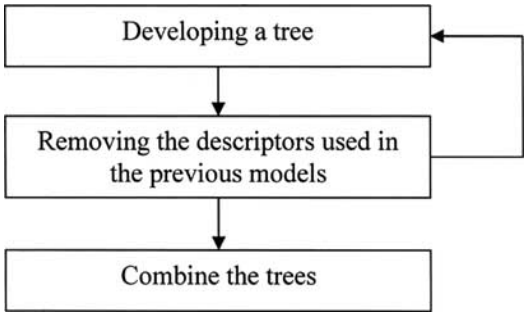


Fig. (3). Overview of Decision Forest (DF). The individual trees are developed sequentially, where each tree uses a distinct set of descriptors. Classification (i.e., prediction) of an unknown chemical is based on the mean results of all trees.

EXTERNAL VALIDATION VERSUS CROSS-VALIDATION

A model fitted to the training set has minimal utility unless it can be generalized to predict unknown chemicals. The ability to generalize the model is an imperious requirement for the regulatory acceptance of QSARs. Most experts in the QSAR field, as well as the present authors, are emphatic that a model's predictive capability minimally needs to be demonstrated using some sort of cross-validation or external validation procedures. Although both procedures share many common features in principle, they are different in both ability and efficiency in assessing a model's overall prediction accuracy, applicability domain and chance correlation during implementation and execution.

When sufficient data is available, a fitted model should be validated by predicting chemicals not used in the training set, but whose activities are known (the test set). This external validation method is analogous to a real-world application. However, external validation lacks validity unless the test set is sufficiently large and diverse, and encompasses the chemical space of the intended application. Using a small number of test set chemicals is inadequate for validation and also possibly wastes valuable data that otherwise could improve the overall quality of a model. Design of the test set in terms of size and diversity and, most importantly, suitably for the intended application is the principal prerequisite for acceptable external validation. Most QSAR models for which external validation has been used does not use a designed test set for a pre-specified application domain. As a matter of fact, it is surely meaningless to design a test set without knowing how the model will be applied, which, in fact, accounts for most QSAR applications. Although the external validation method might be able to discover the classes of chemicals that are not well predicted by the model, it generally provides only an overall assessment of a model with little indication of the prediction confidence for individual chemicals. In other words, *external validation is of little value for assessing the applicability domain.*

A common practice for defining a test set in external validation is to randomly select a portion of chemicals from a dataset. From this perspective, cross-validation provides a similar measure of model performance for a given and fixed set of chemicals. In cross-validation, a fraction of chemicals in the training set are excluded, and then predicted by the model generated from the remaining chemicals. When each chemical is left out one at a time, and the process is repeated for each chemical, it is known as leave-one-out cross-validation. If the training set is divided into *N* groups with approximately equal numbers of chemicals, and the process is repeated for each group, it is called *N*-fold cross-validation. The 10-fold cross-validation procedure is commonly used to assess the predictive capability of a classification model. It appears that, by comparing with external validation, cross-validation provides a systematic measurement of a model's performance without the loss of chemicals set aside for testing. It is necessary to point out that the cross-validation results vary for each run due to random partitioning of the dataset, and thus it is recommended to repeat the cross-validation process many times [32]. The average result of the multiple cross-validation runs provides an unbiased assessment of a model's predictivity.

Most, if not all, classification methods require selection of the relevant or informative descriptors before modeling is actually performed. This is necessary because the method could otherwise be more susceptible to the effects of noise. The a priori selection of descriptors, however, carries with it the additional risk of so-called "selection bias" [33] when the descriptors are selected before the dataset is divided into the training and test sets (Fig. **4A**). Because of selection bias, both external validation and cross-validation could significantly overstate prediction accuracy [34]. To avoid selection bias, the descriptor selection should be made after the dataset splitting (Fig. **4B**). It appears that this procedure is much easier to implement in external validation than in cross-validation because the computational cost maybe prohibitive for iterative descriptor selection during cross-validation for many classification methods. However, the tree-based methods, including both DT and the consensus tree methods, hold the advantage of avoiding the selection bias during cross-validation, because the model is developed at each cycle by selecting the descriptors from the entire pool of descriptors. The cross-validation thereby provides a realistic, unbiased assessment of the predictivity in a consensus tree model.

Given a constant set of chemicals, we feel that cross-validation is more powerful in measuring model performance than external validation with respect to assess overall prediction accuracy, applicability domain and chance correlation.

MODEL VALIDATION

In general terms, SAR model validation has the purpose of demonstrating the overall prediction quality of the model.

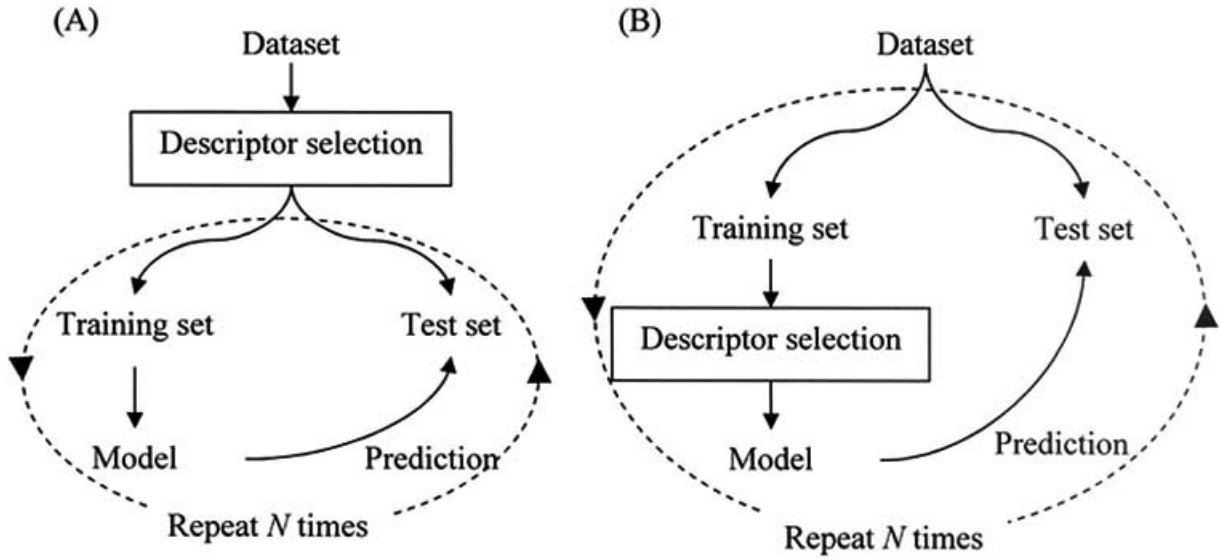


Fig. (4). Two procedures involving descriptor selection in validation processes: (A) descriptor selection occurs before dataset splitting (selection bias); and (B) descriptor selection occurs after dataset splitting (correct procedure). The solid line illustrates the external validation process while both solid and dashed lines together are for the cross-validation process.

How and the extent to which validation is done, however, largely depends on the model's intended use.

If the model is to be applied to a known population of chemicals, regulatory acceptance of the model could depend entirely on the results of validation carried out that is specific to the particular chemical population. The model's validity can be demonstrated by comparing the predicted results with the experimental results on an external test set that is objectively selected from the application population. Consequently, the unbiased selection of an appropriate test set becomes an essential step in determining the validity of the model. The selected chemicals should represent the diversity of molecular structure and activity of the application population, and the selection process should provide statistically significant data to assess false positives and false negatives. Fig. **5** depicts a model-driven selection method to determine a test set that meets the aforementioned criteria.

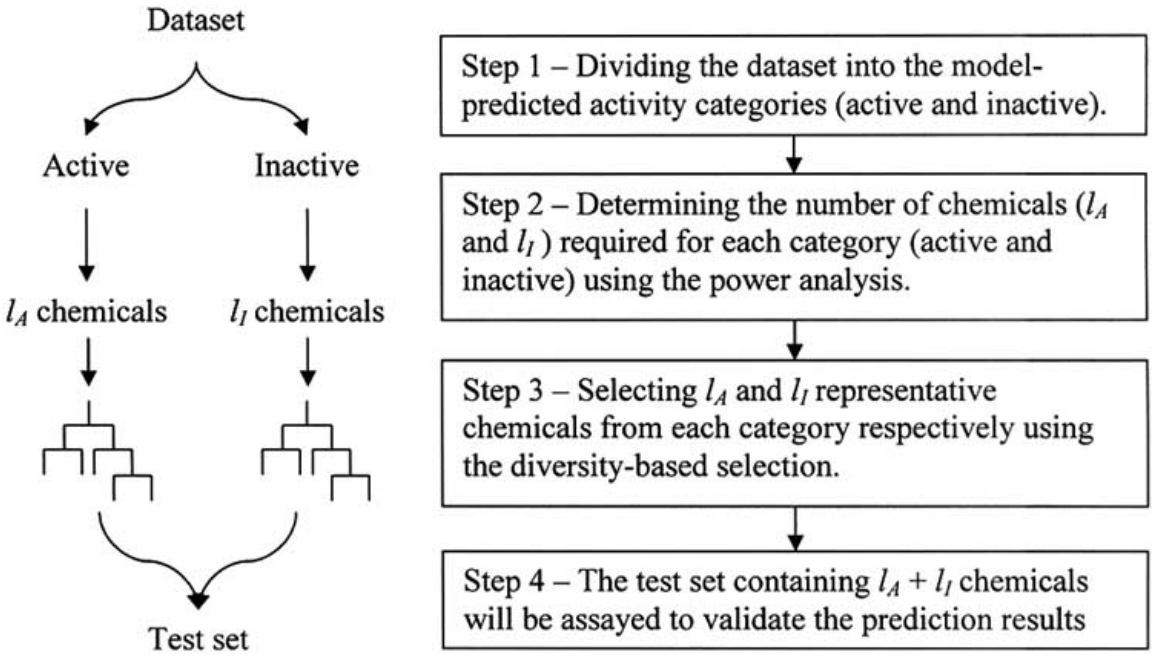


Fig. (5). Schematic presentation of a model-driven selection method to construct a test set. In Step 3, two diversity analysis methods can be used to select *N* chemicals from one category ($N=l_A$ for the active category and $N=l_I$ for the inactive category): (1) Group this category of chemicals into *N* clusters on the basis of their structural similarity using clustering methods, and one chemical from each cluster is randomly selected; and (2) Group this category of chemicals into *n* clusters ($n<N$), different number of chemicals are randomly selected from each cluster using a weighted factor, and the total number of chemicals will be *N*. Both approaches have been used in drug discovery for hit selection, and the weighted approach has proven to be more efficient.

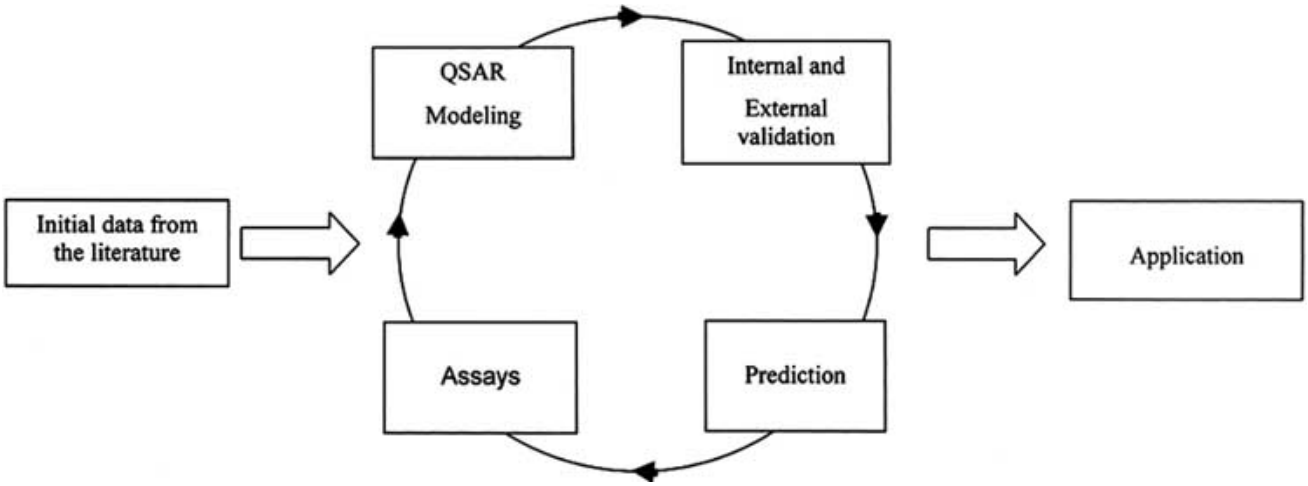


Fig. (6). Depiction of the recursive process used in our lab to develop QSAR models for predicting estrogen receptor binding. The process starts with data from an initial set of chemicals from the literature for QSAR modeling. Next, the preliminary QSAR models are used prospectively to define a set of chemicals that will further improve the model's robustness and predictive capability. The new chemicals are assayed, and these data are then used to challenge and refine the QSAR models. Validation of the model is critical. The process emphasizes the living model concept.

In many cases, however, an intended chemical application domain is broadly or vaguely defined, such as "the model will be used to predict estrogenic activity for the environmental chemicals". In these cases, the chemical structure domain intended for application is not entirely known, and is perhaps not entirely knowable, prior to either model development or validation. Validation for such models is best undertaken as part of a recursive process of incremental improvement, as depicted in Fig. **6**, where such a process was employed to predict potential environmental endocrine disruptors [35]. More specifically, the QSAR model is essentially a "living model" that is successively challenged as new data are available, where incorrectly predicted chemicals are investigated to determine whether their inclusion in the training set will further improve the model's robustness and predictive capability. If so, the new chemicals will be assayed and incorporated into the model. Although this is a natural process with many benefits [12,35], it has the drawback of being *reactive* in nature. While confidence in model prediction may grow over time as the training set expands, there is no quantitative measure of prediction accuracy available when the model is challenged by new untested chemicals.

APPLICABILITY DOMAIN IN CONSENSUS TREE METHODS

Discussion in the literature regarding the accuracy or acceptability of QSARs will often state the need to define a QSAR's applicability domain. In fact, however, applicability domain means different things to different modelers, and would unavoidably be quantified with different parameters for different modeling and statistical techniques [10]. Since a single simple definition of applicability domain is not feasible, and its conceptual purpose is in model quality, applicability domain might best be viewed for measures of confidence in *each prediction* when the overall quality of a model is acceptable.

We describe applicability domain for consensus tree models as being determined by two parameters: (1) *prediction confidence*, or the certainty of a prediction for an unknown chemical; and (2) *domain extrapolation*, or the prediction accuracy of an unknown chemical that lies outside of the chemistry space of the training set chemicals [36]. Both parameters can be quantitatively estimated in the consensus tree approaches, where individual models are developed based on DT. Taken together, prediction confidence and domain extrapolation assess the applicability domain of a model for each prediction.

For each tree in a consensus tree model, the probability (0-1) for an unknown chemical to be active is taken to be the percentage of active chemicals in the terminal node to which the chemical belongs. Consequently, for the consensus tree model, the mean probability value for a chemical can be calculated by simply averaging the probabilities across all individual trees (or other combining methods such as voting). Chemicals that have a probability larger than 0.5 are designated active, whereas those that have a mean probability less than 0.5 are designated inactive. Importantly, this mean probability is also a measure of the confidence of each prediction. Specifically, larger probabilities approaching one correspond to high confidence that the chemical is active. Correspondingly, smaller probabilities approaching zero correspond to high confidence the chemical is inactive. Conversely, probabilities near 0.5 are the most equivocal and correspond to low confidence whether the prediction is active or inactive [37].

Fig. **7** gives an example illustrating how prediction accuracy and prediction confidence are related. Prediction accuracy is plotted versus prediction confidence for both DF and DT for a problem where 2000 runs of 10-fold cross-validation for a dataset (ER232 that contains 232 chemicals) that was used to model estrogen receptor binding activity [36]. A strong trend of increasing accuracy with increasing confidence is readily apparent for both DF and DT, as is the substantially higher accuracy for DF across the entire range

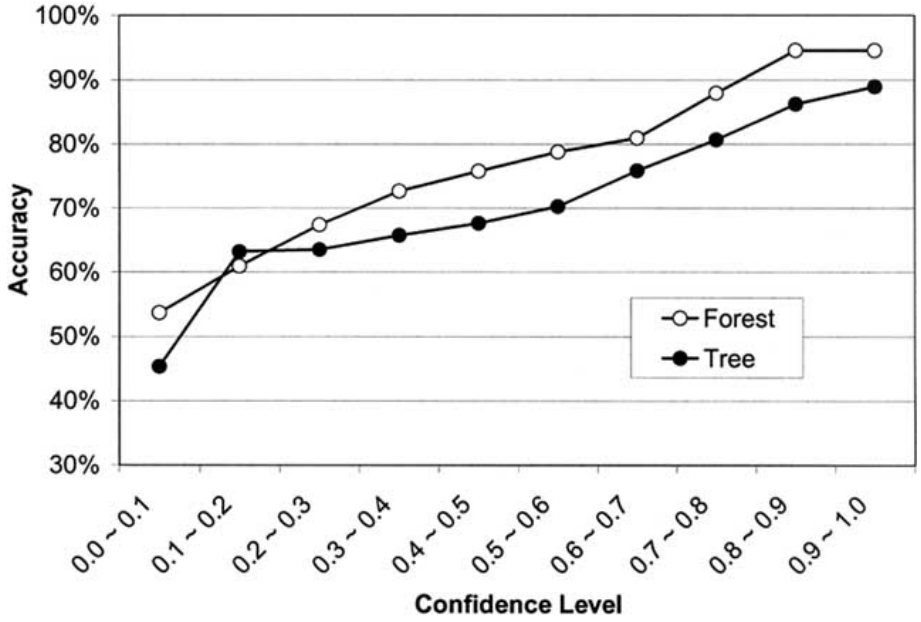


Fig. (7). DF prediction accuracy versus confidence level for ER232 based on 2000 runs of 10-fold cross-validation.

Table 1. Prediction Accuracy in the HC and LC Regions for ER232 Based on 2000 Runs of 10-Fold Cross-Validation

	Decision Forest		Decision Tree	
	High	**Low**	**High**	**Low**
Prediction Accuracy	86.5%	63.8%	77.3%	66.7%
	81.9%		76.7%	

of Confidence Level (The Confidence Level is defined as $|P_i - 0.5| / 0.5$, where P_i is the probability value for chemical i.). Table **1** compares the high confidence (HC), low confidence (LC) and overall prediction accuracy when HC and LC are defined as Confidence Level > 0.4 and < 0.4, respectively. For DF, the HC prediction accuracy is ~86%, ~22% higher than the LC prediction accuracy (~64%), and ~5% higher than the overall prediction accuracy. In contrast for DT, the HC prediction accuracy is only marginally better than total prediction accuracy. The minimum improvement in DT is a direct consequence of over fitting with DT resulting in 95% of predictions being HC, compared with only 80% for DF. The results demonstrate that the consensus tree model gives a better assessment of prediction confidence than does the single tree model.

Defining the training domain is the prerequisite for assessing the domain extrapolation. Commonly in QSAR, a training domain is viewed as an N-dimensional space defined by the N descriptors used by the model and could be called the global domain. In a tree, however, the classification of an unknown chemical is determined by only one terminal node that is descendent from the root node through a set of "IF-THEN" rules based on k descriptors x_i (i=1, … k) as illustrated in Fig. **8**. Analogous to the global domain, the training domain for a tree, and by rational extension for a DF model, can be defined as a k-dimensional space, called the focused domain. For the datasets we have evaluated, there was no appreciable difference in results for global domain compared to focused domain [36]. Fig. **9** shows the results of evaluation of DF domain extrapolation for two training datasets, ER232, and a larger ER1092 (contains 1092 chemicals with estrogen receptor binding activity) [36]. Specifically, the plot compares the overall prediction accuracy for chemicals within the training domain with accuracy for chemicals falling several degrees of

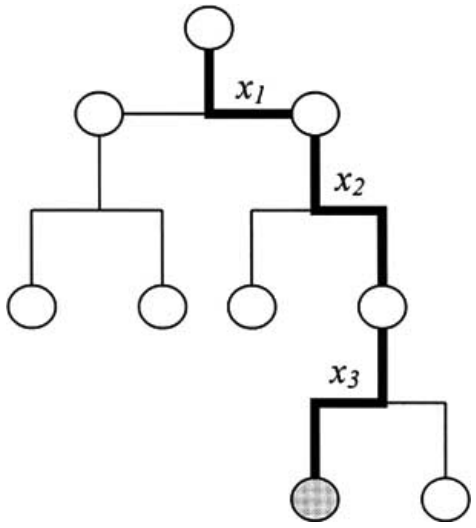


Fig. (8). The focused domain representing the training domain of a tree. For an unknown chemical predicted by the tree, its classification is determined by a terminal node (e.g., dark circle) to which it belongs. There are three descriptors used in the path (bold line) from the root to the terminal node and the range of these three descriptors across all chemicals in the training set determines the training domain.

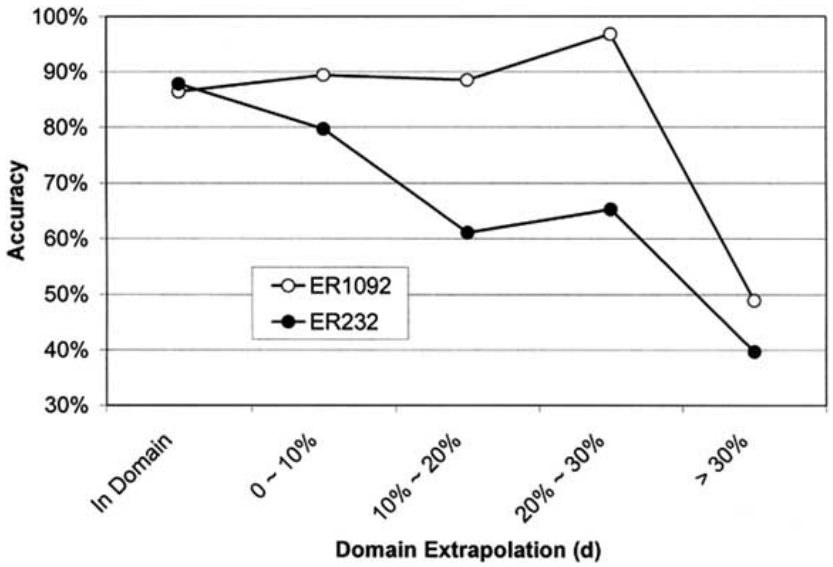
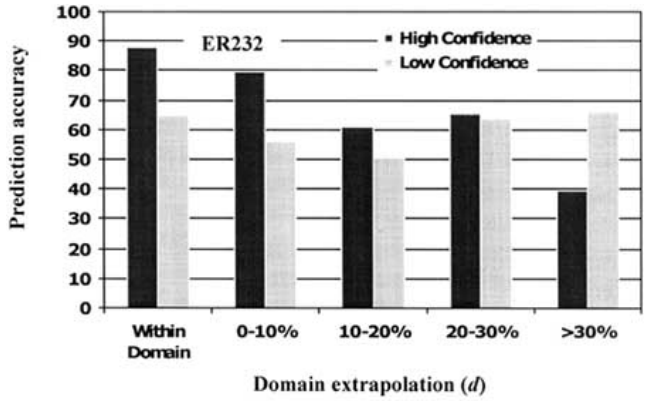


Fig. (9). DF prediction accuracy versus domains extrapolation for ER232 and ER1092 based on 2000 runs of 10-fold cross-validation. Domain extrapolation (*d*) for a chemical is defined as a percentage away from the focused domain as illustrated in Fig. (**8**), while the prediction accuracy for the domain *d* is calculated by dividing correct predictions by total number of chemicals in this domain.

extrapolation outside the focused domain. In general, the farther away the chemicals are from the training domain, the lesser the prediction accuracy. For ER232, the prediction accuracy is some 10% less for chemicals having 10% extrapolation. In contrast, for ER1092, a marked decrease in accuracy only occurs beyond 30% extrapolation.



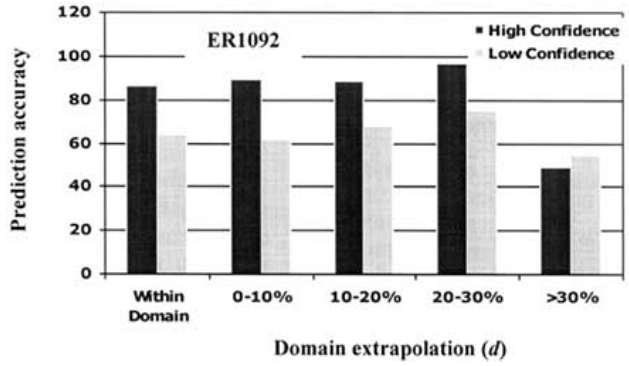


Fig. (10). DF prediction accuracy versus domain extrapolation for ER232 and ER1092 in both HC and LC regions based on 2000 runs of 10-fold cross-validation.

It is generally acknowledged that a model's prediction accuracy largely depends on the size and diversity of the training set. It is also widely acknowledged that models from a larger training set yield predictions with greater confidence, even for the chemicals that are less well represented by the training set. Despite the broad acceptance of these concepts, they have actually not been well tested in a quantitative sense. The results shown in Fig. **7** and **9**, taken together, offer compelling substantiation that larger training sets provide predictions with both greater accuracy and confidence. The results also suggest that a DF model's applicability domain can be well determined by two inextricably linked concepts, prediction confidence and domain extrapolation.

Fig. **10** combines the results of Figures **7** and **9** in order to examine the relative importance of prediction confidence and domain extrapolation. Prediction accuracy is plotted versus domain extrapolation for each dataset (ER1092 and ER232), with separate bars distinguishing predictions in HC and LC domains. Figure 10 shows the trend of decreasing prediction accuracy with increasing extrapolation in the HC region, for both ER232 and ER1092 datasets, which is consistent with the results shown in Figure **7**. In contrast, the LC region prediction accuracy is consistently lower and exhibits no discernable trend with extent of extrapolation for both datasets. The results imply that the model's applicability domain is predominantly determined by prediction confidence. For the HC predictions, the accuracy is dependent on domain extrapolation. However, for the LC predictions, the prediction accuracy is substantially lower and fairly insensitive to extent of domain extrapolation.

ASSESSMENT OF CHANCE CORRELATION

Testing whether a fitted SAR model is, in fact, a chance correlation is highly recommended. Testing becomes increasingly imperative for smaller training data sets, with increasing numbers of descriptors, with increasing noise in biological data, and with an increasing skew of numbers of chemicals across activity categories. All of these conditions increase the omnipresent risk of obtaining a chance correlation lacking predictive value.

To assess the degree of chance correlation, we first generate many pseudo datasets (e.g., 2000 pseudo datasets) using a randomization test, where the activity classification is randomly scrambled across all chemicals in the training set. Next, we apply a 10-fold cross-validation on each of pseudo datasets. The null distribution, i.e., the distribution of prediction accuracy for all pseudo datasets, can then be compared with the distribution of multiple 10-fold cross-validation results derived from the real dataset. The degree of chance correlation in the predictive model can be estimated from the overlap of the two distributions.

Fig. **11** shows the results of a test for chance correlation of DF models to predict liver carcinogenicity based on four

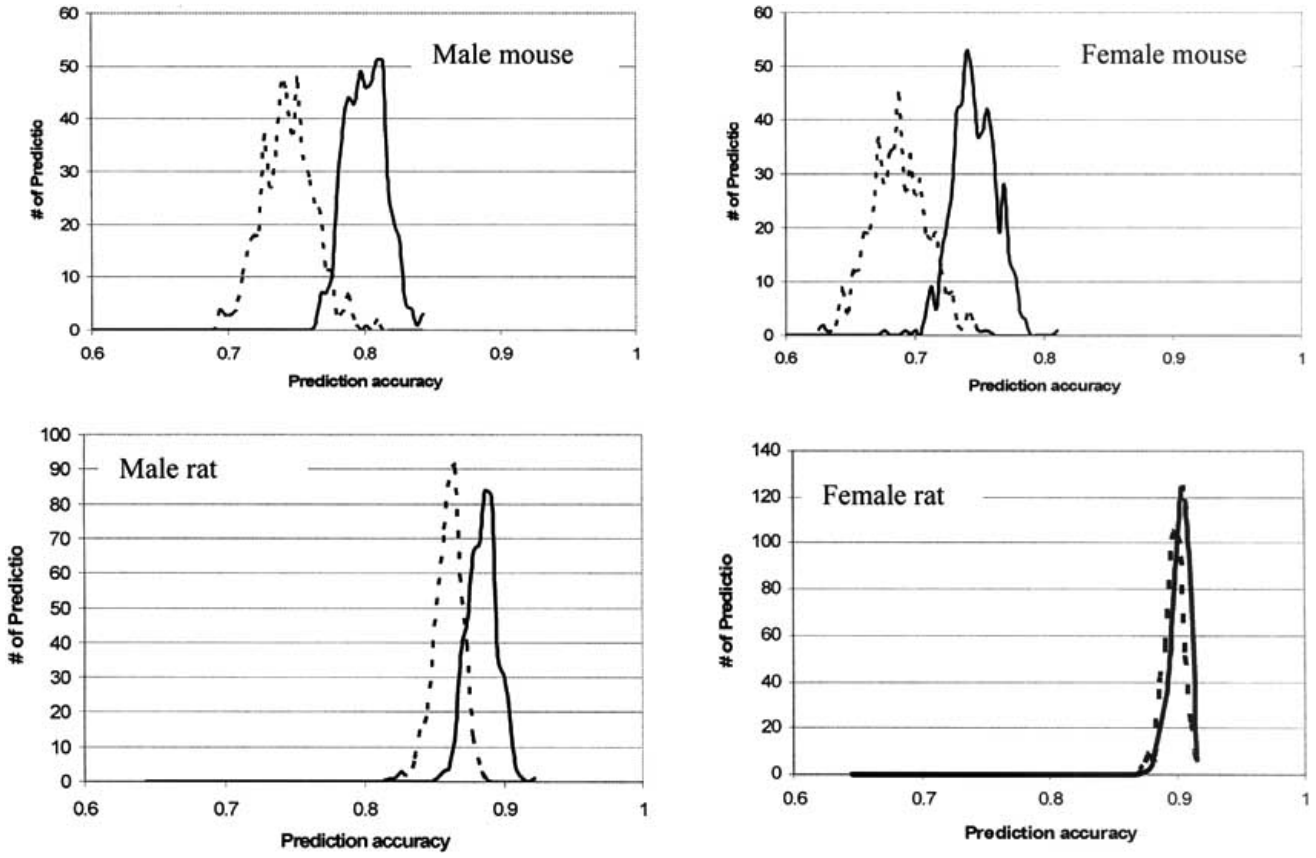


Fig. (11). Assessment of the chance correlation in DF for four datasets listed in Table **3**. For each graph, the null distribution (-------) is generated from the results of 10-fold cross-validation on 2,000 pseudo datasets while the real distribution (——) is derived from 2,000 runs of 10-fold cross-validation for the original dataset.

Table 2. DF 10-Fold Cross-Validation Results for Four Liver Carcinogenicity Datasets Obtained from Testing on Two Species, Rat and Mouse, for both Sexes

Datasets	# of Cmpds	# of carcinogens	# of non-carcinogens	Cross-validation accuracy (%)
Female mouse	247	60	187	74.6
Male mouse	241	48	193	80.1
Male rat	230	28	202	88.5
Female rat	237	21	216	89.8

datasets. Although high cross-validation results were obtained for the models based on these four datasets (Table **2**), there was a significant overlapping between the null and real distribution for each dataset (Fig. **11**), indicating high degree of chance correlations for these models.

In another example, we compared the null distribution for 2,000 pseudo datasets with the real distribution generated from 2,000 runs of 10-fold cross-validation for ER232. As shown in Fig. **12**, the distribution of prediction accuracy of the real dataset centers around 82% while the pseudo datasets are near 50%. The real dataset has a much narrower distribution compared to the pseudo datasets, indicating that the training models generated from the cross-validation procedure for the real dataset give consistent and high prediction accuracy within corresponding test sets. In contrast, as expected, the prediction results of each pair of training and test sets in the 10-fold cross-validation process for the pseudo datasets varied widely, implying a large variability of signal/noise ratio among these training models. Importantly, there is no overlap between two distributions, indicating that a statistically and biologically relevant DF model could be developed using the real dataset.

CONCLUSION

Any QSAR model will produce some degree of error. It is highly desirable in regulatory applications to identify the limitations of a model. Being able to quantitatively assess the accuracy limitation of each specific prediction allows selection of alternative methods, whether *in silico* or experimental, to augment or supplant unreliable predictions, thus improving the value and utility of QSAR-based risk assessment. Assessing model limitations is a vital step towards the regulatory acceptance of QSARs.

A model's limitations should be assessed from three different perspectives: (1) overall model predictivity (model validation); (2) individual prediction confidence (applicability domain); and (3) chance correlation. These can be more readily assessed in the consensus tree modeling such as the DF method than other SAR methods. Using DF as an example, we have found that in DF:

- Combining multiple valid trees that use unique sets of descriptors into a single decision function produces a higher quality model than individual trees.
- The prediction confidence and domain extrapolation can be readily calculated and constitutes the definition of applicability domain for DF.
- Since the descriptor selection and model development are integrated, the cross-validation avoids descriptor selection bias and is a more useful means than the external validation in assessing a model's limitations.
- Carrying out many runs of cross-validation is computationally inexpensive and provides an

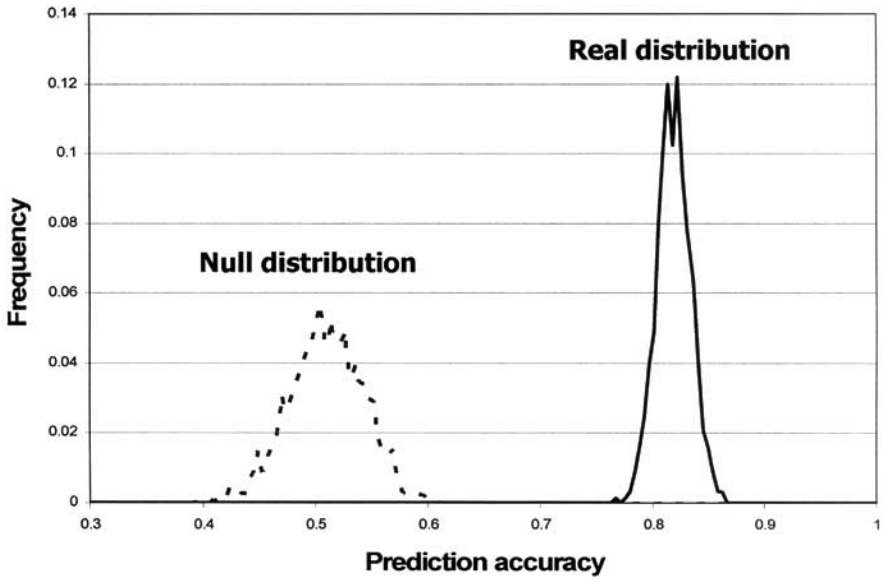


Fig. (12). Assessment of the chance correlation in DF for ER232, an estrogenic dataset that contains 232 chemicals tested in an estrogen receptor binding assay. The same assessment described in Fig. (**10**) was used in this test.

unbiased assessment of a model's predictive capability, applicability domain and potential chance correlation.

REFERENCES

[1] Contrera, J. F.; Matthews, E. J.; Kruhlak, N. L.; Benz, R. D. *Regul. Toxicol. Pharmacol.*, **2004**, *40*, 185-206.

[2] Lessigiarska, I.; Cronin, M. T.; Worth, A. P.; Dearden, J. C.; Netzeva, T. I. *SAR QSAR Environ. Res.*, **2004**, *15*, 169-190.

[3] Walker, J. D.; Carlsen, L.; Hulzebos, E.; Simon-Hettich, B. *SAR QSAR Environ. Res.*, **2002**, *13*, 607-616.

[4] Russom, C. L.; Breton, R. L.; Walker, J. D.; Bradbury, S. P. *Environ. Toxicol. Chem.*, **2003**, *22*, 1810-1821.

[5] Walker, J. D.; Jaworska, J.; Comber, M. H.; Schultz, T. W.; Dearden, J. C. *Environ. Toxicol. Chem.*, **2003**, *22*, 1653-1665.

[6] Tong, W.; Perkins, R.; Fang, H.; Hong, H.; Xie, Q.; Branham, S. W.; Sheehan, D. M.; Anson, J. F. *Regul. Res. Perspect.*, **2002**, *1*, 1-16.

[7] Tong, W.; Fang, H.; Hong, H.; Xie, Q.; Perkins, R.; Anson, J. F.; Sheehan, D. *Pure Appl. Chem.*, **2003**, *75*, 2375-2388.

[8] Cronin, M. T.; Jaworska, J. S.; Walker, J. D.; Comber, M. H.; Watts, C. D.; Worth, A. P. *Environ. Health Perspect.*, **2003**, *111*, 1391-1401.

[9] Jaworska, J. S.; Comber, M.; Auer, C.; Van Leeuwen, C. J. *Environ. Health Perspect.*, **2003**, *111*, 1358-1360.

[10] Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.*, **2003**, *111*, 1361-1375.

[11] Perkins, R.; Fang, H.; Tong, W.; Welsh, W. J. *Environ. Toxicol. Chem.*, **2003**, *22*, 1666-1679.

[12] Tong, W.; Welsh, W. J.; Shi, L.; Fang, H.; Perkins, R. *Environ. Toxicol. Chem.*, **2003**, *22*, 1680-1695.

[13] Bates, J. M.; Granger, C. W. J. *Oper. Res. Quart.*, **1969**, *20*, 451-468.

[14] Bunn, D. W. In *Judgemental Forecasting*; Wiley: New York, 1987; pp 229-241.

[15] Bunn, D. W. *Eur. J. Oper. Res.*, **1988**, *33*, 223-229.

[16] Clemen, R. T. International Journal of Forecasting, **1989**, 5, 559-583.

[17] Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 525-531.

Received: December 7, 2004 **Accepted: December 22, 2004**

[18] Hong, H.; Tong, W.; Fang, H.; Shi, L.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J. D.; Branham, W.; Sheehan, D. M. *Environ. Health Perspect.*, **2002**, *110*, 29-36.

[19] Shi, L.; Tong, W.; Fang, H.; Xie, Q.; Hong, H.; Perkins, R.; Wu, J.; Tu, M.; Blair, R. M.; Branham, W. S.; Waller, C.; Walker, J.; Sheehan, D. M. *SAR QSAR Environ. Res.*, **2002**, *13*, 69-88.

[20] Opitz, D.; Shavlik, J. *Connect. Sci.*, **1996**, *8*, 337-353.

[21] Krogh, A.; Vedelsby, J. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, **1995**; pp 231-238.

[22] Maclin, R.; Shavlik, J. *Proceedings of the Fourteenth International Joint Conference on Intelligence*, **1995**; pp 524-530.

[23] Drucker, H.; Cortes, C. In *Advances in Neural Information Processing Systems*; MIT Press, **1996**; pp 479-485.

[24] Quinlan, J. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*; MIT Press, **1996**; pp 725-730.

[25] Maclin, R.; Opitz, D. *Proceeding of the 14th National Conference on Artificial Intelligence*; American Association for Artificial Intelligence, **1997**; pp 546-551.

[26] Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. *Mutagenesis*, **2004**, *19*, 365-377.

[27] Breiman, L. *Mach. Learn.*, **1996**, *24*, 123-140.

[28] Freund, Y.; Schapire, R. E. In *Machine Learning*: *Proceedings of the Thirteenth International Conference*, **1996**; pp 148-156.

[29] Amit, Y.; Geman, D. *Neural Comput.*, **1997**, *9*, 1545-1588.

[30] Breiman, L.; Department of Statistics, University of California: Berkeley, **1999**.

[31] Gunther, E. C.; Stone, D. J.; Gerwien, R. W.; Bento, P.; Heyes, M. P. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*, 9608-9613.

[32] Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 186-195.

[33] Simon, R.; Radmacher, M. D.; Dobbin, K.; McShane, L. M. *J. Natl. Cancer Inst.*, **2003**, *95*, 14-18.

[34] Ambroise, C.; McLachlan, G. J. *Proc. Natl. Acad. Sci. USA.*, **2002**, *99*, 6562-6566.

[35] Tong, W.; Fang, H.; Hong, H.; Xie, Q.; Perkins, R.; Sheehan, D. In *Predicting Chemical Toxicity and Fate*; CRC press: New York, London, Boca Raton, **2004**.

[36] Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. *Environ. Health Perspect.*, **2004**, *112*, 1249-1254.

[37] Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 525-531.